Rapport d’Analyse Exploratoire Complète

Étude de cas : Analyse des vols aériens

Adama Sall

Date : Avril 2025

1. Introduction

Ce rapport présente l’analyse exploratoire d’un jeu de données issu du secteur aérien, contenant 10 683 observations et 11 variables sur des vols (compagnie, date, source, destination, durée, prix, nombre d’escales, etc.).

2. Description du jeu de données

- **Source :** Fichier Excel « Data_Train.xlsx »
- **Format :** 10 683 lignes × 11 colonnes

Variable	Description	Type
Airline	Compagnie aérienne	Catégorique
Date_of_Journey	Date du vol	Date
Source	Ville de départ	Catégorique
Destination	Ville d’arrivée	Catégorique
Route	Itinéraire	Catégorique
Dep_Time	Heure de départ	Date/Heure
Arrival_Time	Heure d’arrivée	Date/Heure
Duration	Durée du vol (en minutes)	Numérique
Total_Stops	Nombre d’escales	Numérique
Additional_Info	Infos complémentaires	Catégorique
Price	Prix du billet (en monnaie locale)	Numérique

3. Méthodologie adoptée

- **Imputation des valeurs manquantes :** ça dependra du contexte
- **Détection des valeurs aberrantes :** Méthode IQR (intervalle interquartile)
- **Suppression/correction des extrêmes :** Remplacement par les bornes (méthode « bounds »)
- **Standardisation des catégories :** Mapping, homogénéisation des labels
- **Outils :** Python (pandas, seaborn, matplotlib)

4. Inspection initiale

- **Valeurs manquantes :**
  - 1 valeur manquante dans « Route » et « Total_Stops »
- **Doublons :**
  - Aucun doublon détecté après inspection
- **Types de données :**
  - Conversion des dates et heures au format datetime
  - Conversion de « Duration » en minutes (numérique)
  - Mapping de « Total_Stops » en numérique (0 à 4)

5. Correction des erreurs et incohérences

- **Remplacer les valeurs aberrantes par les bornes :**
  - Prix et durée hors bornes corrigés par les bornes IQR
- **Standardisation des catégories :**
  - Mapping des escales : ‘non-stop’ → 0, ‘1 stop’ → 1, etc.
  - Homogénéisation des noms de compagnies, villes, etc.
- **Suppression des lignes avec valeurs manquantes**

  Justification : Le nombre de valeur manquante est très peu par rapport aux nombre d’observation

6. Analyse univariée

6.1 Variables numériques

- **Prix**
    - Min : 1 759, Max : 23 017 (après traitement des outliers)
    - Moyenne : 9 022, Médiane : 8 372, Écart-type : 4 260
    - Distribution asymétrique positivement
- **Durée**
    - Min : 5 minutes, Max : 2 070 minutes
    - Moyenne : 642, Médiane : 520
    - Distribution multiimodale (pics)
- **Total_Stops**
    - Mode : 1 escale (0 ou 1), Max : 2,5 (après traitement des outliers)

Visualisations : Histogrammes, boxplots (avant/après nettoyage)

6.2 Variables catégorielles

- **Airline** : 12 compagnies (Jet Airways la plus fréquente)
- **Source/Destination** : Delhi, Cochin, Banglore dominent
- **Additional_Info** : Majorité 'No info', autres modalités rares regroupées

Visualisations : Diagrammes en barres

7. Traitement des valeurs manquantes

- **Route** et **Total_Stops** : suppression de la seule ligne manquante (impact négligeable)

8. Analyse bivariée

8.1 Corrélations numériques

- **Prix vs Durée** : Corrélation positive modérée ($r \approx 0,45$)
- **Prix vs Total_Stops** : Corrélation faible ($r \approx 0,18$)
- **Durée vs Total_Stops** : Corrélation positive (plus d'escales implique durée plus longue)

Visualisation : Heatmap des corrélations

8.2 Croisements catégoriels

- **Airline vs Prix**
    - Jet Airways et Air India : prix moyens plus élevés

    - GoAir, SpiceJet : prix plus bas
- **Source/Destination vs Prix**
    - Vols vers Cochin et Banglore : prix moyens plus élevés
- **Total_Stops vs Prix**
    - Vols directs (‘non-stop’) : prix généralement plus élevés

8.3 Analyse bivariée avancée (tests statistiques)

- **ANOVA** sur le prix selon la compagnie : différence significative ($p < 0,01$)
- **Test de Kruskal-Wallis** sur le prix selon le nombre d’escales : différence significative ($p < 0,01$)

9. Checklist des étapes

Étape	Statut
Chargement des données	✅
Inspection initiale	✅
Suppression des doublons	✅
Correction des types	✅
Standardisation des catégories	✅
Traitement des valeurs manquantes	✅
Détection des valeurs aberrantes	✅
Suppression/correction des extrêmes	✅
Analyse univariée	✅
Analyse bivariée	✅
Visualisations	✅

10. Visualisations clés

- Histogrammes des prix, durées, escales
- Boxplots (avant/après nettoyage)
- Diagrammes en barres (compagnies, destinations)

- Heatmap des corrélations

11. Conclusions

- **Nettoyage efficace :**
    - Jeu de données prêt pour la modélisation (pas de valeurs manquantes, outliers traités, catégories homogènes)
- **Tendances majeures :**
    - Les prix varient fortement selon la compagnie, la destination et le nombre d'escales
    - Les vols directs sont plus chers mais plus rapides
    - Jet Airways domine le marché en fréquence et leur prix sont plus cher parmi les compagnies.
    - Matin et après-midi sont les moments les plus fréquents pour les départs.
    - Les vols plus longs tendent à coûter plus cher.
    - Le nombre d'escales influence le prix mais moins fortement.
    - Les vols avec plus d'escales sont généralement plus longs.
    - La durée influence le prix mais moi fortement
    - Les nombres de vol sont moins fréquent en Avril et les billets sont moins cher, tandis qu'en mars , mai et juin ils sont chers.
    - La disponiblité ou non des informations sur les vols n'impactent pas les prix
- **Préparation à la modélisation :**
    - Les premières corrélations suggèrent que la compagnie, la durée et le nombre d'escales sont des facteurs explicatifs du prix

12. Recommandations

- Augmenter la fréquence des vols directs avril
- Réviser les prix des vols directs
- proposer des vols en dehors des pics matinaux/après-midi pour répartir la demande et réduire les coûts opérationnels

13. Annexes

- **Code Python complet (Notebook)**

14. Synthèse

L’analyse exploratoire a permis de nettoyer, structurer et comprendre en profondeur le jeu de données aérien. Les premiers résultats montrent une forte influence de la compagnie, du nombre d’escales et de la destination sur le prix des billets.